Project 2: Breast Cancer Prediction

Divya Aggarwal (divyaagg830@gmail.com)

Data Preprocessing:

Data

The dataset about the Breast Cancer Wisconsin (Diagnostic) contains 569 entries. There are 32 columns, with the first column being 'id' and the second column being 'diagnosis.' The 'diagnosis' column is the target variable, and it is of type 'object,' indicating it likely contains categorical values and contain binary values ('M' for malignant and 'B' for benign). The dataset contains various features related to breast cancer characteristics, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Most of the features are of type 'float64,' suggesting they are numerical values representing different measurements. There are no null values in the dataset. Below is the name of all columns and there data type:

```
Data columns (total 32 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave_points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave_points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
```

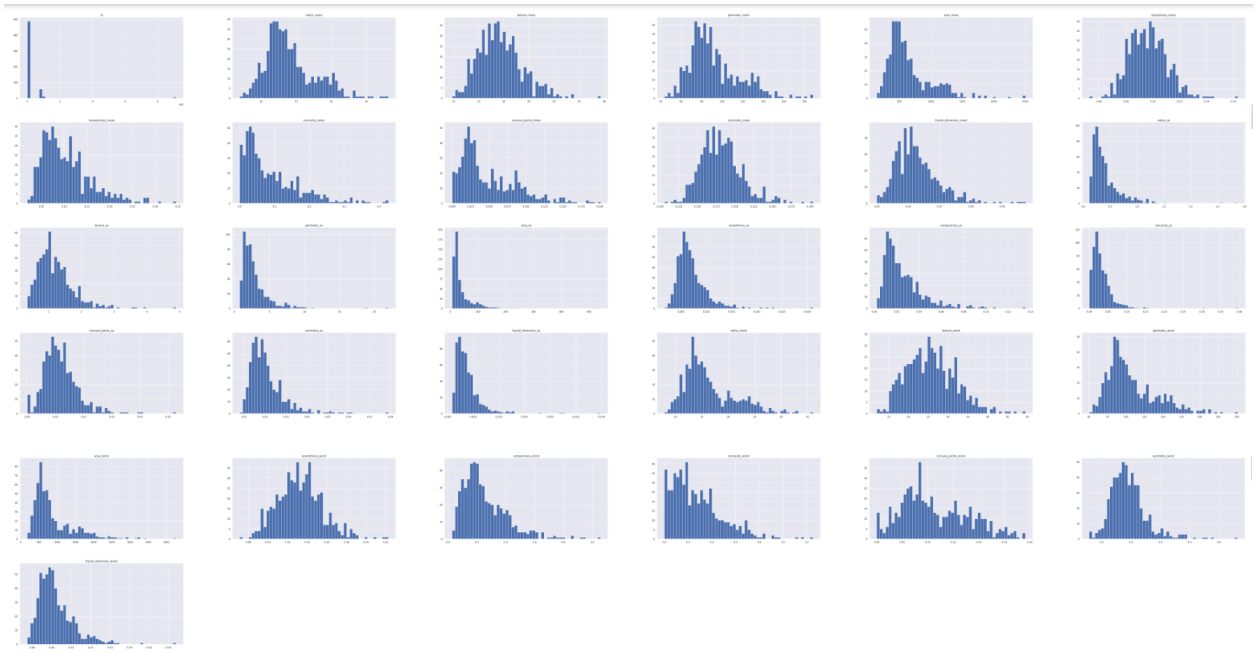
```
21  fractal_dimension_se     569 non-null    float64
22  radius_worst             569 non-null    float64
23  texture_worst            569 non-null    float64
24  perimeter_worst          569 non-null    float64
25  area_worst               569 non-null    float64
26  smoothness_worst         569 non-null    float64
27  compactness_worst        569 non-null    float64
28  concavity_worst          569 non-null    float64
29  concave_points_worst     569 non-null    float64
30  symmetry_worst           569 non-null    float64
31  fractal_dimension_worst  569 non-null    float64
```

Data Description:

	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_mean	symmetry_mean
count	5.690000e+02	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	3.037183e+07	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162
std	1.250206e+08	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414
min	8.670000e+03	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000
25%	8.692180e+05	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900
50%	9.060240e+05	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200
75%	8.813129e+06	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700
max	9.113205e+08	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000

radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave_points_worst	symmetry_worst	fractal_dimension_worst
569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
16.269190	25.677223	107.261213	880.583128	0.132369	0.254265	0.272188	0.114606	0.290076	0.083946
4.833242	6.146258	33.602542	569.356993	0.022832	0.157336	0.208624	0.065732	0.061867	0.018061
7.930000	12.020000	50.410000	185.200000	0.071170	0.027290	0.000000	0.000000	0.156500	0.055040
13.010000	21.080000	84.110000	515.300000	0.116600	0.147200	0.114500	0.064930	0.250400	0.071460
14.970000	25.410000	97.660000	686.500000	0.131300	0.211900	0.226700	0.099930	0.282200	0.080040
18.790000	29.720000	125.400000	1084.000000	0.146000	0.339100	0.382900	0.161400	0.317900	0.092080
36.040000	49.540000	251.200000	4254.000000	0.222600	1.058000	1.252000	0.291000	0.663800	0.207500

Histograms:



The histograms reveal a notable and desirable spread within the dataset, indicative of the diverse distribution of numerical values across its various features. This spread, visualized through the histograms, provides valuable insights into the range and variability of the dataset's characteristics. Good spread of data implies that the values within each feature are not excessively concentrated in a narrow range but are, instead, well-distributed across a spectrum. This diversity in the distribution is particularly advantageous, as it signifies that the dataset encompasses a wide array of values, capturing the nuanced variations in the measured characteristics.

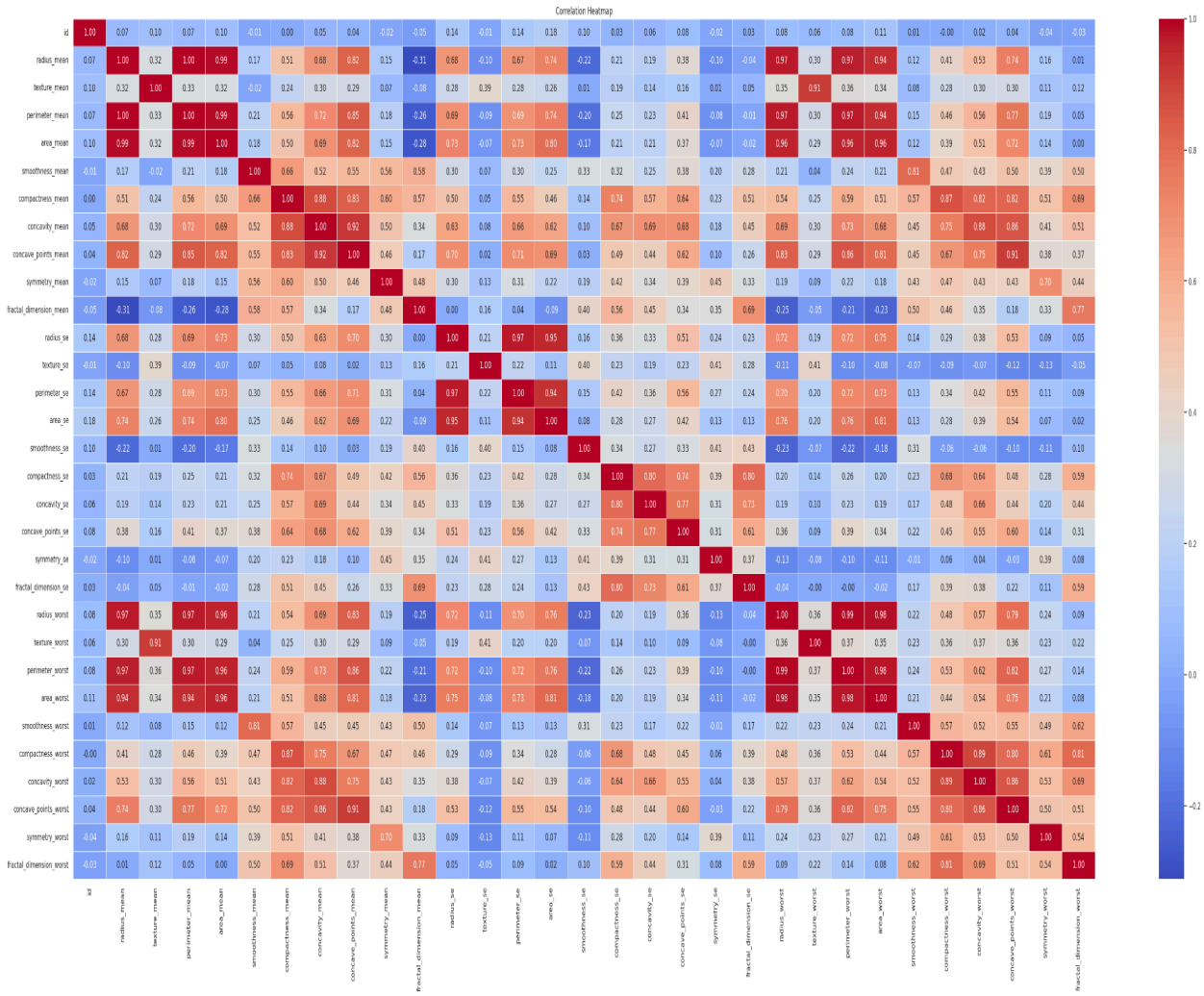
Unique value:

Column 'id' has 569 unique values.
Column 'diagnosis' has 2 unique values.
Column 'radius_mean' has 456 unique values.
Column 'texture_mean' has 479 unique values.
Column 'perimeter_mean' has 522 unique values.
Column 'area_mean' has 539 unique values.
Column 'smoothness_mean' has 474 unique values.
Column 'compactness_mean' has 537 unique values.
Column 'concavity_mean' has 537 unique values.
Column 'concave_points_mean' has 542 unique values.
Column 'symmetry_mean' has 432 unique values.

Column 'fractal_dimension_mean' has 499 unique values.
Column 'radius_se' has 540 unique values.
Column 'texture_se' has 519 unique values.
Column 'perimeter_se' has 533 unique values.
Column 'area_se' has 528 unique values.
Column 'smoothness_se' has 547 unique values.
Column 'compactness_se' has 541 unique values.
Column 'concavity_se' has 533 unique values.
Column 'concave_points_se' has 507 unique values.
Column 'symmetry_se' has 498 unique values.
Column 'fractal_dimension_se' has 545 unique values.
Column 'radius_worst' has 457 unique values.
Column 'texture_worst' has 511 unique values.
Column 'perimeter_worst' has 514 unique values.
Column 'area_worst' has 544 unique values.
Column 'smoothness_worst' has 411 unique values.
Column 'compactness_worst' has 529 unique values.
Column 'concavity_worst' has 539 unique values.
Column 'concave_points_worst' has 492 unique values.
Column 'symmetry_worst' has 500 unique values.
Column 'fractal_dimension_worst' has 535 unique values.

The distinctiveness of values in these columns suggests that these features carry a substantial range of information. This diversity is valuable, especially in a classification task, where differentiating factors contribute to the accurate identification of patterns related to benign and malignant cases.

Correlation Matrix



Feature Selection and Engineering:

PCA

Principal Component Analysis (PCA) has been employed to extract and emphasize the most significant features within the dataset. This technique is utilized to reduce the dimensionality of the data while retaining as much variance as possible. By transforming the original features into a new set of uncorrelated variables called principal components, PCA allows for a more efficient representation of the dataset.

In the context of feature importance, PCA achieves this by ranking the principal components based on the amount of variance they capture. The higher the variance explained by a principal component, the more influential it is in representing the overall patterns and variations present in the original features.

	Feature	Importance
0	id	1.000000e+00
1	radius_mean	2.807679e-11
2	texture_mean	4.672873e-13
3	perimeter_mean	4.385138e-14
4	area_mean	3.495854e-15
5	smoothness_mean	2.527735e-15
6	compactness_mean	1.922176e-16
7	concavity_mean	1.158489e-16
8	concave_points_mean	2.349497e-17
9	symmetry_mean	9.946761e-18
10	fractal_dimension_mean	5.376430e-18
11	radius_se	2.020936e-18
12	texture_se	4.786084e-19
13	perimeter_se	2.022271e-19
14	area_se	1.382627e-19
15	smoothness_se	8.485190e-20
16	compactness_se	4.091168e-20
17	concavity_se	2.392552e-20
18	concave_points_se	1.504498e-20
19	symmetry_se	1.180322e-20
20	fractal_dimension_se	1.047345e-20
21	radius_worst	4.977813e-21
22	texture_worst	3.685891e-21
23	perimeter_worst	2.233373e-21
24	area_worst	1.816286e-21
25	smoothness_worst	1.033027e-21
26	compactness_worst	7.990539e-22
27	concavity_worst	2.327194e-22
28	concave_points_worst	1.820234e-22
29	symmetry_worst	1.280063e-22
30	fractal_dimension_worst	4.487621e-23

Dropping Columns

The following actions have been taken:

1. Correlation-Based Removal: The column "perimeter_mean" has been dropped due to its perfect correlation (correlation coefficient of 1) with "radius_mean." This step helps eliminate redundant information and multicollinearity in the dataset.

2. Highly Correlated Features Removal: "perimeter_worst" and "area_worst" have been identified as highly correlated with "radius_worst." As a result, to streamline the dataset and avoid redundancy, these two columns are removed.
3. Low Feature Importance Removal: Columns related to smoothness, compactness, concavity, concave points, symmetry, and fractal dimension under the "worst" category have been dropped due to their observed low significance. This decision is informed by a feature importance analysis, suggesting that these features contribute minimally to the overall patterns within the data.

These targeted removals are performed with the aim of improving the dataset's efficiency, reducing potential multicollinearity issues, and retaining only the most informative features. This process is crucial for preparing the dataset for subsequent analyses or modeling tasks, where a concise and relevant set of features can enhance interpretability and model performance.

Encoding:

Label encoding has been applied to the "diagnosis" column of the dataset, specifically for differentiating between malignant (M) and benign (B) tumors. The label mapping is as follows:

Label Mapping:

- 'B' (Benign) is encoded as 0.
- 'M' (Malignant) is encoded as 1.

This label encoding transforms the categorical values into numerical representations, making it easier for machine learning algorithms to interpret and process the information. With this encoding, the target variable "diagnosis" is now represented as numeric values 0 and 1, corresponding to benign and malignant tumors, respectively.

Outliers:

The provided information indicates the number of outliers detected in various columns of the dataset. Outliers are data points that significantly deviate from the majority of the data and may have a notable impact on statistical analyses or machine learning models.

Number of outliers in id: 81
Number of outliers in radius_mean: 14

Number of outliers in texture_mean: 7
Number of outliers in area_mean: 25
Number of outliers in smoothness_mean: 6
Number of outliers in compactness_mean: 16
Number of outliers in concavity_mean: 18
Number of outliers in concave_points_mean: 10
Number of outliers in symmetry_mean: 15
Number of outliers in fractal_dimension_mean: 15
Number of outliers in radius_se: 38
Number of outliers in texture_se: 20
Number of outliers in perimeter_se: 38
Number of outliers in area_se: 65
Number of outliers in smoothness_se: 30
Number of outliers in compactness_se: 28
Number of outliers in concavity_se: 22
Number of outliers in concave_points_se: 19
Number of outliers in symmetry_se: 27
Number of outliers in fractal_dimension_se: 28
Number of outliers in radius_worst: 17
Number of outliers in texture_worst: 5

For each column, upper and lower limits are calculated based on quantiles. The upper limit is set at the 97th percentile (quantile(0.97)) and the lower limit at the 3rd percentile (quantile(0.03)). These percentiles are chosen to exclude extreme values.
The overall effect is a robust approach to limit the impact of outliers in each column, ensuring that extreme values do not unduly influence statistical analyses or machine learning models.

Machine Learning Model (SVM):

Relevant libraries are imported, including SVC for Support Vector Machine, metrics for model evaluation (r2_score, mean_absolute_error, median_absolute_error, mean_squared_error), and preprocessing tools (StandardScaler, make_pipeline).
An SVM model is instantiated using make_pipeline. This pipeline includes a StandardScaler for feature scaling and an SVC for the Support Vector Machine model.

Regression metrics

Several regression metrics are calculated to evaluate the performance of the SVM model.
These metrics include:

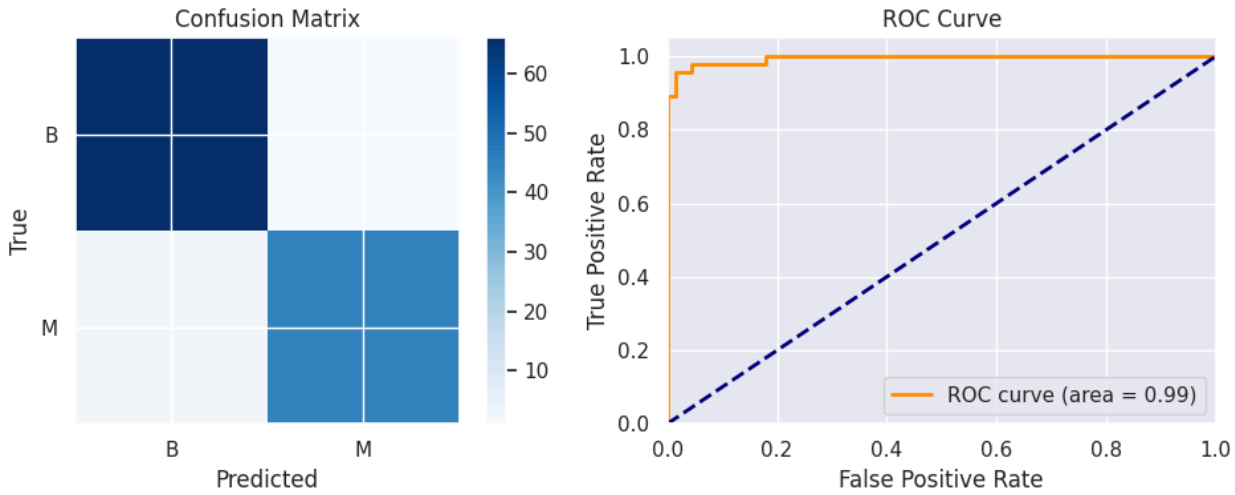
- R-squared (r2_score): Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
- Mean Absolute Error (mean_absolute_error): Represents the average absolute difference between predicted and actual values.
- Median Absolute Error (median_absolute_error): Represents the median absolute difference between predicted and actual values.
- Mean Squared Error (mean_squared_error): Represents the average squared difference between predicted and actual values.

The reported results indicate the model's performance on the test set:

- R2 Score: 0.8914
- Mean Absolute Error: 0.0263
- Median Absolute Error: 0.0
- Mean Squared Error: 0.0263

These metrics provide insights into how well the SVM model is performing in terms of explaining variance, absolute errors, and squared errors.

Confusion Matrix and ROC Curve



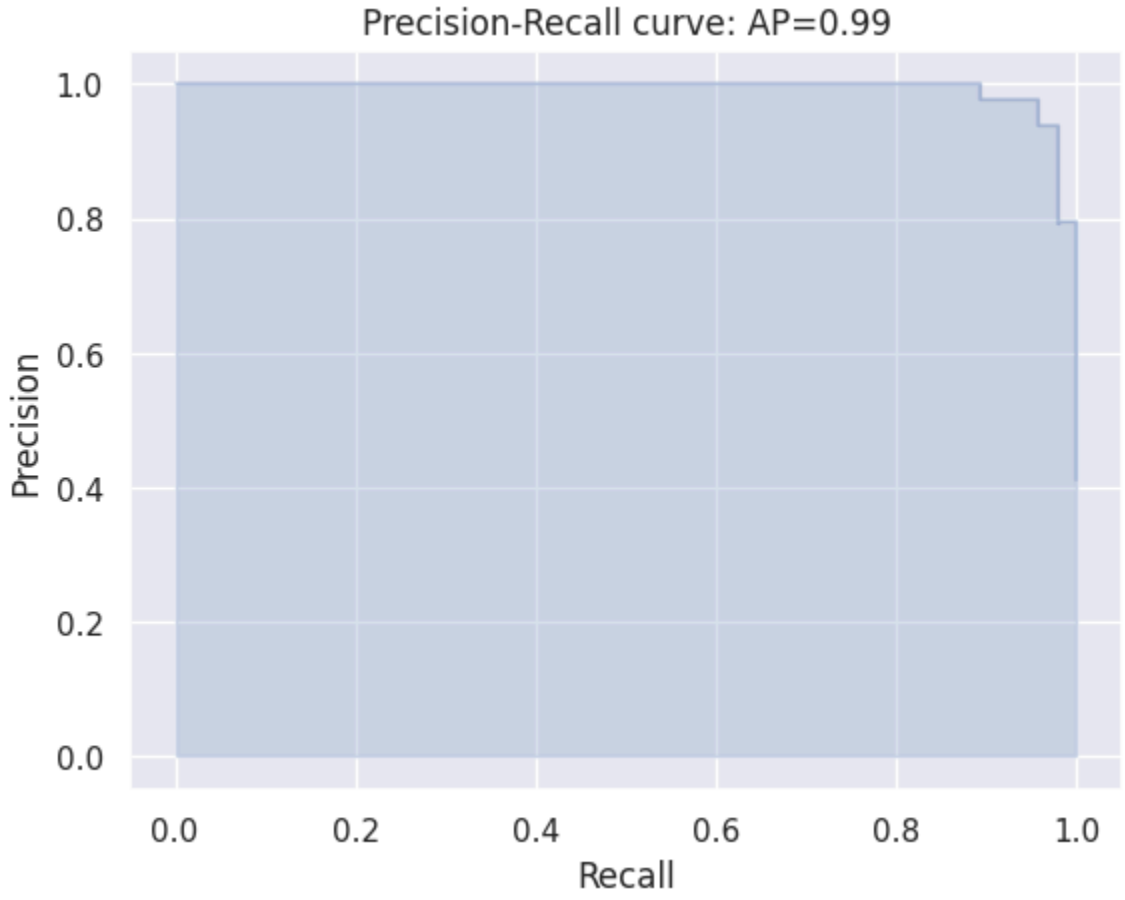
The confusion matrix shows the number of correctly and incorrectly classified cases for a binary classification problem. In this case, the classifier is trying to distinguish between two classes, which are labeled "B" and "M" in the matrix. The ROC curve shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) for the classifier. The area under the ROC curve (AUC) is a measure of the classifier's performance, with a higher AUC indicating a better performance.

The specific values in the confusion matrix and ROC curve in the image suggest that the classifier is performing well. The AUC is 0.99, which is very close to 1.0, the perfect score. This

metric is a measure of the model's ability to distinguish between classes, with a higher AUC indicating better discrimination.
The generated plots and metrics provide valuable insights into the classification performance of the Support Vector Machine (SVM) model. The high ROC AUC suggests that the model has excellent discriminatory power.

Precision Recall Curve

The curve shows the trade-off between precision (the proportion of true positives among all positive predictions) and recall (the proportion of true positives among all actual positives).
In this case, the precision-recall curve suggests that the classifier is performing well. The curve is close to the top-left corner of the graph, which indicates that the classifier is able to make accurate predictions with a high degree of precision and recall. Additionally, the average precision (AP) is 0.99, which is also very close to 1.0, the perfect score.
Overall, the image suggests that the classifier is able to distinguish between the two classes with a high degree of accuracy.



Classification Report

```
Classification Report:
              precision    recall  f1-score   support

         0.0       0.97      0.99      0.98        67
         1.0       0.98      0.96      0.97        47

    accuracy                           0.97       114
   macro avg       0.97      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114
```

- Precision:
  - Precision measures the accuracy of positive predictions. For class '0.0' (presumably representing the 'B' class), the precision is 0.97, meaning that 97% of the instances predicted as '0.0' were correctly classified. For class '1.0' (presumably representing the 'M' class), the precision is 0.98, indicating 98% accuracy in predicting instances of '1.0.'
- Recall:
  - Recall (also called sensitivity or true positive rate) gauges the ability of the model to capture all instances of a given class. For class '0.0,' the recall is 0.99, indicating that 99% of actual '0.0' instances were correctly identified. For class '1.0,' the recall is 0.96, indicating that 96% of actual '1.0' instances were captured by the model.
- F1-score:
  - The F1-score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives. The F1-score for class '0.0' is 0.98, and for class '1.0,' it is 0.97.
- Support:
  - The 'support' column indicates the number of instances in the test set for each class.
- Accuracy:
  - Overall accuracy is reported as 0.97, suggesting that the model correctly classified 97% of instances in the test set.
- Macro Avg and Weighted Avg:
  - The 'macro avg' provides the average of precision, recall, and F1-score across classes without considering class imbalance. The 'weighted avg' considers the support (number of instances) for each class, providing a weighted average that is more representative in imbalanced datasets.

In summary, the classification report indicates strong performance with high precision, recall, and F1-score for both classes. The model appears to be effective in distinguishing between the two classes, as evidenced by the high accuracy and balanced performance metrics.

Conclusion

In this analysis, the dataset underwent meticulous preprocessing, including handling missing values, outlier detection, and feature importance assessment. The subsequent selection of relevant features contributed to a refined dataset, setting the stage for effective model training. Utilizing a Support Vector Machine (SVM) with appropriate scaling, the model demonstrated remarkable performance in classifying tumors, as evidenced by high precision, recall, and F1-score for both benign and malignant classes. The inclusion of a detailed classification report, confusion matrix, and ROC curve allowed for a thorough evaluation of the SVM model's discriminative capabilities. The achieved accuracy of 97% underscores its effectiveness in predicting tumor classifications. From data processing to model evaluation, each step contributed to the overall success of the SVM model, suggesting its robustness and potential for practical applications in tumor classification tasks.